Acoustic Measures for Real-Time Voice Coaching

Ying Li
Giving Tech Labs
Seattle, United States
ying@giving.tech

Abraham Miller*
Giving Tech Labs
Seattle, United States
abe.m@giving.tech

Arthur Liu*
Giving Tech Labs
Seattle, United States
arthur.l@giving.tech

Kyle Coburn
Giving Tech Labs
Seattle, United States
kyle.c@giving.tech

Luis J. Salazar
Giving Tech Labs
Seattle, United States
luis@giving.tech

ABSTRACT

Our voices can convey many different types of thoughts and intent; how our voices carry them is often not consciously controlled and as a consequence, unintended effects may arise that negatively impact our relationships. How we say things is as important as what we say. This paper presents methodologies for computing a set of physical properties from sound waves of a speaker's voice directly, referred to as acoustic measures. Experiments are designed and conducted to establish the correlations between physical properties and auditory measures for human perception of sound waves. Based on these correlations, a voice coaching app can guide users, in real-time or deferred retrospective, to modify their speech's auditory measures, such as rate of speech, energy level, and intonation, to achieve their intended communication goals.

CCS CONCEPTS

• Human-centered computing → Mobile computing; **• Applied computing → Sound and music computing**; **• Hardware → Sound-based input / output**; **• Information systems → Data mining**.

KEYWORDS

voice analysis, speech signal processing, mobile application

ACM Reference Format:
Ying Li, Abraham Miller, Arthur Liu, Kyle Coburn, and Luis J. Salazar. 2020. Acoustic Measures for Real-Time Voice Coaching. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20), August 23–27, 2020, Virtual Event, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394486.3403326

1 INTRODUCTION

Our voices can convey many different types of thoughts and intent; in many cases, we do not consciously control how our voices carry them and as a consequence, we fail to achieve the communication results we wanted, and unintended effects may arise that negatively impact our relationships. To effectively communicate, *how* we say things is as important as *what* we say.

The challenges in voice communication are often amplified in stressful situations such as public speaking, class presentations, business meetings, emergency situations, or conversations related to healthcare, among others. These challenges extend to caregivers of persons on the spectrum of neurological disorders such as Autism, Down's syndrome, and others, when their voices convey unintended emotions to persons under their care, causing, in extreme cases, sensory overload in the listener.

While there is extensive scientific and popular research related to the ideal level of auditory properties – inflection, rate of speech, tone, volume, modulation, and other parameters – or the inferred attributes such as the speaker's emotional state, there is not an easy way for an individual to assess these auditory properties and inferred attributes in real-time. Even after a verbal communication session, it is hard to objectively reflect and improve the way in which the individual communicates with others.

Of particular importance is the preservation of individual privacy. Most existing technologies used for voice coaching rely on speech-to-text and subsequent analysis of the transcribed text. Other technologies rely on the analysis of video recordings for facial and body language expressions. These are intrusive and in some applications that involve interactions with patients or with underage individuals, existing privacy laws would limit or prohibit the use of speech to text technologies. Furthermore, rich speech information is lost in the conversion; sarcastic speech, for example, often *means* the opposite of what is literally *said.*

The purpose of the work presented in this paper is to help individuals improve the way they communicate. We compute the acoustic properties of sound waves of a speaker's voice, and present them in terms of auditory measures such as the speaking rate, inflection, and energy level. The computation of these acoustic properties is built into a smart phone app that enables both live and deferred feedback so the speaker can take early intervention while speaking, and improve their speech over time. The analysis methods and applications are patent pending by Giving Tech Labs.

Our contributions presented in this paper are:

(1) a set of mathematical analysis of sound waves of a speaker's voice, without the use of speech-to-text, to calculate voice

[*] Work performed as *AI for Public Interest (AI4PI) Research Fellow* at Giving Tech Labs.

KDD '20, August 23–27, 2020, Virtual Event, CA, USA

ACM ISBN 978-1-4503-7998-4/20/08…$15.00
https://doi.org/10.1145/3394486.3403326

characteristics such as speed of speech, inflection, and energy level

(2) design and execution of tests to certify the correctness and robustness of the implementation

(3) an app that incorporate these computations on smart phones for providing real-time and deferred voice-coaching to individuals

The rest of this paper is organized as follows: Section 2 provides an overview on topics relevant to our application of voice signal analysis and auditory perception of speech. Section 3 introduces the acoustic measures used and our methods for computing them. Section 4 presents the design of experiments and reports on the results. Section 5 introduces the application we built for smart phones. Lastly, Section 6 concludes and outlines our plan for future work.

2 DOMAIN SUBJECT BACKGROUND

The domain subject of the applied data science work presented in this paper belongs to the area of voice signal prosodic analysis. Specifically, we focus on the acoustic measures that can be computed objectively and directly from a sound. Acoustic measures are distinguished from auditory measures that describe human perception of a sound. The two types of measures are highly correlated, as outlined in Table 1 (taken from [17] page 30 with our additions).

Table 1: Correlations between acoustic (physical) measures and auditory (perceptual) measures of voice signal.

acoustic measures	auditory measures
fundamental frequency (denoted as F0)	pitch
intensity	loudness
spectral characteristics	timbre
onset/offset time	timing, speaking rate
phase difference in binaural hearing	location
variation of measures over time units	intonation

We should note that the correlations between acoustic measures and auditory measures as listed in Table 1 are not linear, nor one-to-one. There exist many more acoustic properties that correlate to auditory perceptions, often multiple acoustic features are used together as features in machine learning systems for inferring different auditory properties. These have been researched in the fields of signal processing and spoken language understanding.

For the purpose of this paper, we focus on the computation of the acoustic measures in Table 1 to serve a Voice Coaching Application. For voice coaching, acoustic measures need to be computed in real time from the voice signal directly. Then correlations with auditory measures need to be established, based on which, the computed acoustic measures can be compared against a reference range of auditory measures. We describe our methodologies and experiments on computing the acoustic measures and determining the correlations with auditory measures. For instance, to coach on speaking rate, duration of speech units are computed, mapped to speaking rate, and compared to "ideal speaking rate" for the intended speaking effect.

2.1 Auditory Measures of Human Voices

Auditory measures are studied as part of intonation system of a language [16]. Desired speaking styles differ for the audience groups, the speaking context (presentation vs. one-on-one dialog), speaking distance (large public presentation vs. small group discussion), speaker demographics, and so on. The relationships between these factors, the impact they have on how a voice signal is perceived, and the intended effect of the speech are complex, some studies treat one individual auditory measure and some impact factors while others conduct comprehensive studies on more factors. For instance, Sorokowski et. al. [36] studies relations between speakers' modulation of their speaking fundamental frequencies and the perceived authority on the subject matter the speakers covey. Rodero [32] analyzed the impact of radio announcers' speaking rate has on the perception of subjective assessment in the news. Yuan et al. [39] studies the relation between speaking rate, speaker demographics, and speaking topics in conversations.

2.2 Acoustic Measures from Voice Signals

Acoustic measures and their underlining computational features from voice signals have been utilized for detecting and classifying various properties, such as speaker personality traits (age, gender, personality) [28, 34, 35]), speaker state (affection, intoxication, stress) [2, 7, 25], acoustic events carried within a voice clip [38], emotions carried in real-life conversations [8], and so on.

Acoustic measures have been researched and studied in the fields of signal processing [30], speech recognition [17], computational linguistics, etc., for many decades and still gain active research attention today as the number of new application scenarios and voice devices grows even more abundantly. For example, research and development for accurate estimation of fundamental frequency (F0) dates back to the 90's [3] and earlier, and continue to be topics of research interests [5, 27].

Tools and software libraries for estimating acoustic measures have also been available [4] for decades and continue to emerge [24]. However, our application scenarios for voice coaching require the computation of the acoustic measures on a smart phone or watch without pre-trained models. As such, we need to implement real time estimation of acoustic measures, including our own version of F0 estimator in C, even though F0 has been studied for years by researchers.

2.3 Rate of Speech

Algorithms for estimating rate of speech are utilized in speech recognition systems and motivated by the observation in the mid 90's that strong correlations exist between the performance of speech recognition systems and deviation of test data from the average rate of speech in training data [26]. As such many algorithms make estimation during or after some recognition process, usually interwoven with the speech recognition process and utilizing some learner to detect the speech units. Other algorithms rely on training data that are phonetically transcribed, such as TIMIT [13]. A third types of algorithms usually operate on the amplitude or spectral characteristics and their moments or rate of change of various mathematical properties.

The measure of speaking rate often used in the general public is "words per minute". However, this measure is flawed in a few ways. Firstly, it has coarse granularity, making it unstable for short speech samples. Also, it produces different results depending on the vernacular, dialect, or language used – speech containing many longer words could have low wpm even if spoken quickly. Finally, it requires speech-to-text processing, which is invasive of privacy, and potentially poses limitations in cases of serving the needs of teachers, caregivers, and parents.

Finer speech units are syllables and phonemes. The syllable is a well defined phonological unit in a given language, however, listeners may not agree on the number of syllables heard. Furthermore, not all syllables will be phonetically realized in faster speech than in slow articulation. Our focus in this paper is on counting the *physically observable phonetic units* (phonemes) over time as a proxy for the *heard* speaking rate.

Previous research has shown that there is a significant correlation between the maximum spectral transition positions and the manually selected phoneme boundaries [12]. This, in turn, suggests that there is an important relation between the commonly accepted phoneme boundaries and the perceptual critical points [10]. An application for helping stuttering patients was built based on the computation spectrum transition maximum [1]. We will present our implementation of spectrum transition measure augmented with voice activity detection in Section 3.1.

2.4 Voice Activity Detection

Voice Activity Detection (VAD) algorithms determine what parts of an audio signal contain human speech. Speech analytic systems employ VAD to help improve analysis accuracy, and thanks to the progress in speech recognition, there are many VAD algorithms, among which, RNN VADs are lauded as the most accurate. However, other simpler 'baseline' VADs tend to work nearly as well in signal conditions without large amounts of noise [37]. VAD for real-time applications needs to be effective while being lightweight and robust.

WebRTC [29] is an open-source project that provides robust and efficient protocols for real time audio and video communication in web browsers and mobile applications. Alongside its APIs for communication protocols, WebRTC provides generally useful algorithms for audio and video processing, including a VAD. WebRTC VAD is implemented with a Gaussian Mixture model, and is commonly used as a starting point for speech analysis applications.

3 ACOUSTIC PROCESSING AND MEASUREMENT

The steps to compute acoustic measures are depicted in Figure 1. The input to the processing and computation is a sound file in a standard audio file format recorded on conventional devices. Off-line data used for testing and experimentation consists of benchmark data sets (we describe two in Section 4). On-line data for real-time computation is recorded with the standard recording functionalities on mobile devices. The output consists of aggregated acoustic measures that are not personally identifiable. The following acoustic processing and measures are presented in detail in this section:

- voice activity detection (VAD)
- spectral transition measure (STM)
- signal intensity/energy
- variations of acoustic measures including F0, zero-crossing rate (ZCR)

The set up of our audio signal processing are described in Figure 2. Here, we follow a typical process of performing computations on 32 ms frames of the audio values. We then compute on overlapping frames as we look at frames shifted by 10 ms intervals [17, 30]. This overlap is necessary as it ensures that we do not lose information at the edges of frames.

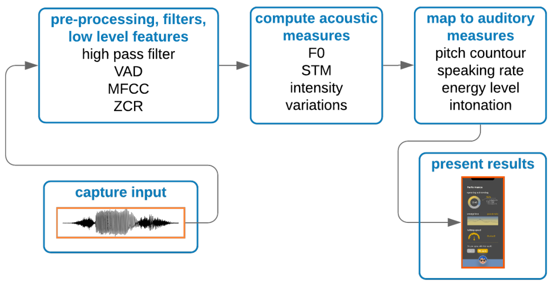


Figure 1: Processing steps for computing acoustic measures.

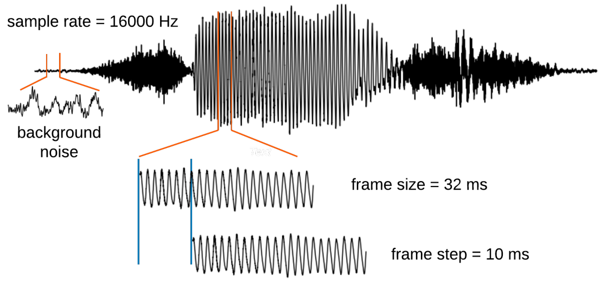


Figure 2: Audio signal processing frame set up.

Before computing any acoustic measures on speech, we first apply a VAD to parse out what parts of an audio file or stream to actually analyze. We choose WebRTC VAD to keep complexity low for initial release. However, after applying the WebRTC VAD to live user audio, we observe that, although its performance is good in identifying the beginning and end of spoken segments, it is too sensitive to quiet noise at the beginning of detection before it adapts to the recording environment, often producing false positives during these quiet unspoken sections.

Thus, we augment the WebRTC VAD to remove from the analysis any audio segments which are below a reasonable intensity level for speech. The modified WebRTC VAD provides effective speech detection on our iPhone recording hardware. Figure 3 depicts an example of VAD result.

3.1 Speech Rate with Spectral Transition Measure (STM)

Since we want to estimate phoneme count through analyzing spectral properties of voice signal, we implemented the spectral transition measure as outlined in [10]. It can be interpreted as the

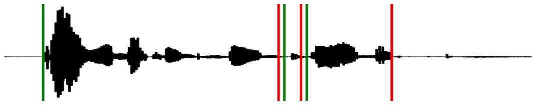


Figure 3: VAD example, green and red depict the beginning and end of voice chunks respectively.

magnitude of the spectral rate of change:

$$STM(m) = \sum_{i=1}^{D} a_i^2(m)/D$$

where D is the dimension of the spectral feature vector, and a_i is the regression coefficient, or the rate of change of the *MFCC*:

$$a_i(m) = \frac{\sum_{n=-I}^{I} MFCC_i(m+n) * n}{\sum_{n=-I}^{I} n^2}$$

where I represents the number of frames on each side of the current frame used in the computation of regression coefficients. Our computation uses $I = 2$ for a 10ms frame step.

MFCCs are meant to approximately represent the human auditory system's response, and are commonly used as features in speech recognition [17].

Once we compute the STM for each frame, phoneme boundary is detected based on STM in relation to the STM of adjacent frames. Peaks in STM value are candidates for phoneme boundaries, and they are identified by selecting frames which have STM value higher than both of neighbors. These peaks are then filtered down to those that are most likely to be phone boundaries by a proprietary method. The remaining filtered STM peaks can be considered to be equivalent to phoneme boundaries.

Using the count of estimated phoneme boundaries computed for a segment of speech, we find the speaking rate of that segment by dividing the amount of computed phonemes by the duration in seconds of the speech segment. This leaves us with a speech rate measured in phonemes per second.

Since our application is exposed to audio input that contains segments of both user speech as well as background noise, we collect STM peaks only in segments that are determined as containing voice by our VAD. Likewise the duration of speech used in the denominator of speaking rate is pared down based on spoken segments found in the audio, while leaving in short unspoken segments which naturally comprise some phonemes; this is discussed further in (Section 4.2).

3.2 Loudness

An important voice coaching topic is controlling the intensity of speech in order to not speak too loudly or too softly for a given context. Often, *loudness* is used to refer to the auditory measure for volume of a sound, while *intensity* of a sound refers to an acoustic measure of the power carried by a sound wave.

There are several acoustic measures of sound intensity. One is Root Mean Square Amplitude (RMS), defined as the square root of the sum of the squared magnitudes of a signal divided by the duration of the signal:

$$RMS(x) = (\frac{1}{N}\sum_{0}^{N} x(t)^2)^{\frac{1}{2}}$$

RMS is a good initial candidate for approximating loudness because of its simplicity and clear relationship to the perception of spoken volume. After conversion to user-legible scale (dB), it provides a rough but usable proxy for speaker loudness [11]. This is our current implementation in the beta version of the coaching application, although there are major limitations to this approach. We now discuss several issues which inform our future work in improving loudness estimation in our application.

First, the subjective perception of a loud or quiet sound is not easily derivable from the physical signal intensity. Converting intensity to a dB scale is a rough approximation, but there are many other methods of measuring loudness of a sound, and care should be taken to measure loudness in a way which is most suited to a specific application [23].

Secondly, there is the distance dependence problem. The recorded intensity of a sound will greatly change depending on the distance between the source and the recording device. With our current scheme, users of our application are instructed to place their phone a particular distance from their mouth for every session. This can be resolved by finding speaker distance first [14], and then computing intensity at the source according to the inverse square law to achieve consistent intensity readings regardless of how far a speaker is from the phone.

Finally, there is the problem of unknown microphone hardware and preprocessing of audio input. The intensity of the digital signal may not correspond well to the physical sound intensity due to unknowns in the recording workflow. To get past this issue, experiments must be done for every hardware setup we plan on releasing to, comparing the results between the app calculation and a physical sound pressure level meter.

Since our current implementation in the app for measuring speaker loudness is temporary with known issues, we leave it as an area of future improvement and do not include experiments for validation like with the other voice analysis features discussed.

3.3 Intonation with Variations in Acoustic Measures

Besides the speed and energy of speech, there remains a great wealth of interesting suprasegmental features with rich implications. In making our application, we required some measure of tonal variation to coach users to speak less monotonously or help them reign back excessively polytonous speech. We call this speech feature *intonation*. The challenge of intonation is providing a quantitative measurement of this characteristic. In our search, we decided on straightforward criteria for such a measurement: (1) it must be low when evaluated on boring, monotonous speech, (2) it needs to increase when the speech is more tonal, and (3) it needs to be high when presented with exciting and lively speech.

While zero-crossing-rate (ZCR) is typically used in voice activity detection implementations [9], if we isolate only the voiced portions and look at the behavior of ZCR, we see that it provides meaningful acoustic information that can distinguish the two emotional categories. (See Figure 4). This suggests that it is a good candidate measurement for intonation which we experiment with later.

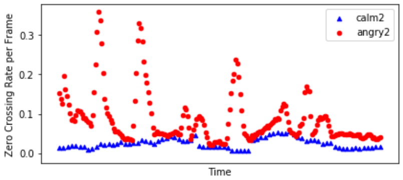


Figure 4: Comparison of Zero-Crossing-Rate for Intense Anger (anger2) and calm (calm2) on voiced portions

In terms of calculation, our ZCR for a length T, substituted in as our frame size, is computed as follows:

$$zcr = \frac{1}{T-1}\sum_{t=1}^{T-1} 1_{\mathbb{R}<0}(s_t s_{t-1})$$

Fundamental frequency (F0) is another candidate metric for intonation we will experiment with, whose relationship with tonality is more intuitive both. We estimate F0 by maximizing the lag for autocorrelation to determine the wave's period, where the correlation for lag k is:

$$r_k = \frac{\sum_{i=1}^{N-k}(Y_i - \bar{Y})(Y_{i+k} - \bar{Y})}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

Through experimentation in Section 4.4, we find that variance and interquartile range of both ZCR and F0 are insightful measurements that satisfy our criteria for a quantitative measurement of intonation.

4 EXPERIMENTATION

4.1 Datasets for Experimentation

The TIMIT Speech Corpus [13] provides phonemically and lexically transcribed speech of American English speakers of different sexes and dialects. Each transcribed element has been delineated in time. It contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. Its text prompts consist of phonetically-compact sentences and phonetically-diverse sentences. The corpus is a very popular benchmark test set for phone recognition [31]. We mainly utilize the time delineated phonemes for testing our computation of speaking rate.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [22] is a validated multimodal database of emotional speech and song. It consists of 7356 clips, gender balanced with 24 professional actors, each performing 104 unique vocalizations with emotions of calm, happy, sad, angry, fearful, surprise, and disgust, at two levels of emotional intensity, normal and strong. (Notice that the neutral emotion has only one emotional intensity.) Researchers use RAVDESS for emotion recognition from audio-visual data. We use the audio signal and the emotion tagging for intonation testing.

4.2 Duration of Intraspeech Gaps

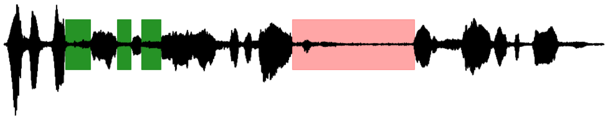


Figure 5: Intraspeech gaps (green) and interspeech pauses (red)

In order to calculate speaking speed, we take the number of phonemes spoken and divide by the amount of time taken for speaking. The STM algorithm (Section 3.1) gives us a solid estimation for the number of phonemes spoken in a speech segment. However, finding the total duration of speech to use as the denominator is more nuanced. There are occasional large noticeable pauses in speech that should not be included in the total duration, there are also natural gaps that occur between phonemes as part of words or sentences. We refer to them as *interspeech pauses* and *intraspeech gaps*, respectively. If an audio input is masked by VAD to determine the total spoken time, then we would underestimate spoken time by removing the small intraspeech gaps. On the other hand if we leave everything in then we will count large interspeech pauses and vastly overestimate.

This experiment is to determine the statistical attributes of normal intraspeech gaps so we can distinguish them from the larger interspeech gaps. We apply our VAD to TIMIT data set to chunk up all the recordings into voiced and unvoiced sections. After chunking up our signals into spoken and unvoiced sections, we discard all the unvoiced segments that were at the very start or end of the recordings. Then, all remaining unvoiced sections are intraspeech gaps. Statistical analysis on the time durations of the intraspeech gaps then should help us in distinguishing kinds of gaps in speech.

Of the 6300 TIMIT audio files, 30708 intraspeech unvoiced segments were detected and analyzed. Our experiment found that the mean duration of these segments was **0.0489 seconds**, and **99.7% of intraspeech gaps are shorter than 0.221 seconds**. For the purpose of determining speaking time in the voice-coaching app, we can count all speech gaps shorter than 0.221 seconds as being part of continuous speech.

4.3 Robustness of STM as Phoneme count estimator

Verifying the accuracy of our STM calculation in estimating how many phonemes are spoken in a given clip will allow us to have confidence in both the VAD which undergirds it, and the estimation for rate of speech we built on top of it. This experiment is meant to both find a constant conversion between calculated STM peak count and phoneme count, and illustrate the strength of this relationship.

We compare the results of our STM peak counting calculation to the known phonemes count in a given speech segment. From the literature, we expect a one-to-one ratio of STM to phoneme [10]. We seek show to show a similar correlation between our STM result and known phoneme amount in a given sentence. To test the strength of the correlation and simultaneously find a constant to convert between STM count and phoneme count, a linear regression is performed using individual TIMIT clips as datapoints along the two dimensions of STM count and phoneme count.

TIMIT has a variety of different speakers available which is good for testing the robustness of our STM results across changes in Speaker voice. We use a modified TIMIT corpus for our experiment that has each audio file padded to the same length with equivalent background noise in order to isolate STM count as the sole independent variable. We also want to validate our STM results across different recording environments with different amounts of noise. To achieve this, we make additional versions of our padded TIMIT with different amounts and types of noise superimposed on the audio, following the example of existing voice analysis evaluation research [6]. We used both white noise and noise collected from our office at signal to noise ratios (SNR) between 20db to -10db.

A linear regression on the padded TIMIT dataset yielded a conversion of **0.9210 STM peaks per phoneme**, with an adjusted r^2 value of **0.840** for this model. Compared to the theoretical one-to-one ratio of STM peaks to phonemes, our calculated STM peak count appears to be a bit lower than expected. Some STM peaks are not found, which is to be expected from our conservative methodology of only classifying peaks if they are certainly phoneme boundaries.

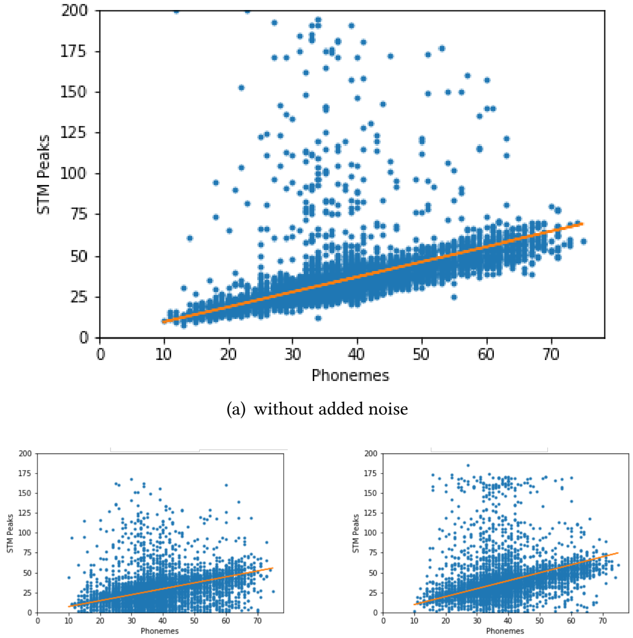
Also worth mentioning is that there are a handful of outlier datapoints ($n \approx 120/6300$) that have much higher STM peak counts than the rest of the data; this effect is illustrated in Figure 6(a). These outliers likely arise from recordings which have a particular kind of background noise that triggers our VAD for large sections even when participants are not speaking, for instance someone quietly talking in the background of a recording. Such an effect would cause our algorithm to severely overcount STM peaks throughout the clip. Despite the outlier clips, this experiment still validates the STM as being highly correlated with phoneme count with a ratio near one-to-one.

How well does the correlation hold up in environments with different amounts of noise? In a white noise environment, we see a large drop off in correlation strength with a signal to noise ratio (SNR) of less than 5db, as seen by a reduction in standard error, RMSE, and adjusted r^2 value. In the simulated office noise environment, the drop off comes earlier with poor results coming from a 10db SNR (Table 2). Additionally, in office noise, which has office babble scattered throughout, we see more of the outliers like described above, while white noise reduces precision of the regression uniformly. Both effects are illustrated in Figure 6(c) and 6(b).

For now, we can safely use our STM peak count to approximate phoneme count within medium-high (>10db) SNR environments in pursuit of estimating rate of speech for an audio clip, relying on our empirically determined conversion of 0.9210 STM peaks per phoneme. According to test users, this parameter produces qualitatively good results in measuring rate of speech on the application.

4.4 Correlations between Acoustic Variations and Intonation

To find a way of computationally estimating speech intonation, we must find a correlation between one or more candidate measures (discussed in 3.3) and the auditory perception of intonation. However, there is no speech recording dataset labeled with intonation levels. We get around this by taking emotionality of speech as a proxy for intonation, since highly emotional speech should have more intonation, and different spoken emotions intuitively have different intonation levels. The RAVDESS dataset is analyzed in this experiment; it has multiple categories and intensities of emotional speech.



(a) without added noise

(b) white noise, 5db SNR

(c) office noise, 15db SNR

Figure 6: STM:phoneme regression plots

First, we map the emotional categories onto an intonation scale. By definition, the neutral clips in the dataset are a good baseline for boring, monotonous speech. Some emotions, such as anger and happiness, are decidedly more tonal than our baseline, but other categories such as calm or sadness are less apparent. By listening to the clips, as well as envisioning the common expressions of these emotions in speech, we placed neutral, calm1, calm2, and sad1 all in the lower intonation bracket, while the remaining emotion categories we consider substantially tonal. Additionally, we expect that all the level two versions of emotions will be higher on the intonation scale than the lower intensity variants, e.g. anger2 will have higher intonation than anger1.

Note that sad1 is the only emotional group not considered substantially tonal, even while sad2 is included. This is because we found sad1 tends to manifest as despondent and low-energy speech, but the more intense version of sad involves the fluctuations in tone associated with sobbing.

To determine if our candidate intonation measures reflect our hypothesized emotion groupings, we computed the Kruskal-Wallis H test [20] for each candidate measure, comparing the neutral category to each of the other emotion groups. Kruskal-Wallis tests for stochastic dominance; the existence of stochastic dominance would illustrate that this other emotion should *not* be grouped together with monotonous, neutral speech, and a lack-thereof would corroborate that it could be grouped together with this category. Of course, there is no proscribed definition for intonation, so our different candidate metrics for intonation may not correspond perfectly to the subjective perception of it. However, a clearly errant grouping would indicate that this metric is not usable.

Looking first at the interquartile-range (IQR) of ZCR in Table 3, comparing neutral with both calm emotions and the weaker version of sad results in H values lower than 6.635 (in bold), which corresponds to an alpha value of 0.01 with one degree of freedom (d.f.). This indicates that we cannot reject H0 that neither group stochastically dominates the other. Thus we can infer that a low-intonation group exists with neutral, calm, and sad1.

Looking at the rest of the results comparing neutral with the other emotions, each of the tests result in a much higher H-value

Table 2: STM:Phoneme regression statistics for different amounts and types of added noise

	No added noise	20db SNR		15db SNR		10db SNR		5db SNR		0db SNR		-5db SNR		-10db SNR	
	—	white	office	white	office	white	office	white	office	white	office	white	office	white	office
Phoneme Coef.	0.9210	1.0180	0.9774	1.7810 *	0.9976	0.7146	1.1677	0.7443	1.5643	0.8079	1.8449	1.4934	3.2413	2.6489	3.6732
adjusted r^2	0.840	0.802	0.849	0.639 *	0.740	0.691	0.612	0.724	0.568	0.504	0.595	0.488	0.819	0.652	0.888
Standard Error	0.005	0.006	0.005	0.017 *	0.007	0.006	0.012	0.006	0.017	0.010	0.019	0.019	0.019	0.024	0.016
RMSE	16.02	20.12	16.40	53.35 *	23.53	19.04	37.05	18.31	54.376	31.94	60.658	60.94	60.737	77.05	51.88
Skew	6.523	3.694	4.149	0.882 *	3.216	2.289	1.913	2.356	0.841	2.863	0.449	0.803	-0.911	-0.381	-1.054

* Further inspection is needed for this setting. It produces results out of step with the general trend and may have been run with the wrong parameters.

that rejects H0, indicating that they can be grouped separately as high-intonation (emotional) speech. Altogether, the low and high intonation split matches our earlier hypothesis.

Applying further analysis with a Welch (unequal variance) T-Test to compare these two groups, it indicates that the mean ZCR IQR for high-intonation is 0.0133 higher than that of low-intonation, with t=-20.95 and p=1.84e-84. This difference is captured well by Figure 7, where the grayed boxes are members of the low-intonation group.

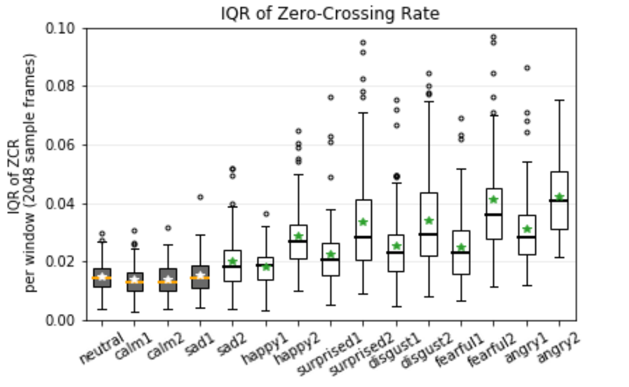


Figure 7: Comparison of Zero-Crossing-Rate Interquartile-Range for Emotion Categories (Shaded are low-intonation group)

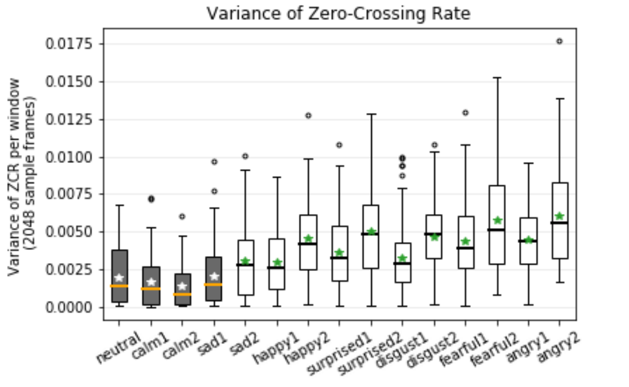


Figure 8: Comparison of Zero-Crossing-Rate Variance for Emotion Categories (Shaded are low-intonation group)

Figure 7 also demonstrates how the IQR of ZCR heuristic satisfies another criteria for our intonation measurement. By comparing the IQR of ZCR within each of the emotions (eg. angry 1 vs. angry 2, excluding the "non-emotional" categories neutral and calm), we see that the more intense version of the emotion results in a larger IQR. This indicates that our heuristic successfully captures the influx of intonation when a speaker increases the display of their emotions; a measurement of the IQR of ZCR acts as a useful proxy for the intonation level of speech. From the Kruksal results in Table 3, we can verify that these mean differences are statistically significant by referencing the H-values from the Kruskal test for each emotional pair. Here, the H-values are all easily above 6.635 (for alpha=0.01, d.f.=1), corroborating our conclusions.

Statistical variance instead of IQR yields very similar results. Using the same procedure, we plot the results to decide our groups and apply Kruskal-Wallis to decide a low-intonation vs. high-intonation grouping. Again, these results gave us the same groups, but this time the grouping is not as robust. For IQR, all pair-wise Kruskal comparisons in the low-energy group yield high p-values, but for variance of ZCR, Calm2 against Neutral narrowly fails to reject H0. From Figure 8, we can see that calm2 has lower variance measurements. As a measure of intonation, this is valid as enforcing calmness in speech can result in significantly more monotony than neutral speech, explaining a decreased H-value (indicating different distributions) when applying Kruskal. Once again though, the comparison of low-energy to high-energy emotion satisfies the criteria that the lower-intonation group results in lower quantitative mean.

Additionally, variance measurements also satisfy the criteria for comparisons between varying intensities of the same emotion as visible in 8. From these results, we can conclude that both IQR and Variance of Zero-Crossing Rate are powerful measurements to quantify the intonation of speech.

Two of the other metrics that we tested were IQR and variance of F0. See Figure 9. For both of these, Kruskal-Wallis resulted in a worse group clustering than ZCR, as it could not as confidently distinguish fearful and disgust from neutral for both IQR and variance. Despite this, we still see a trend otherwise reminiscent of that described in our criteria. Low emotions remain on the low spectrum and many pairwise comparisons within emotions are successful. Overall, variation of F0 may be useful as an intonation measure in conjunction with ZCR, but alone its results do not strongly satisfy our criteria.

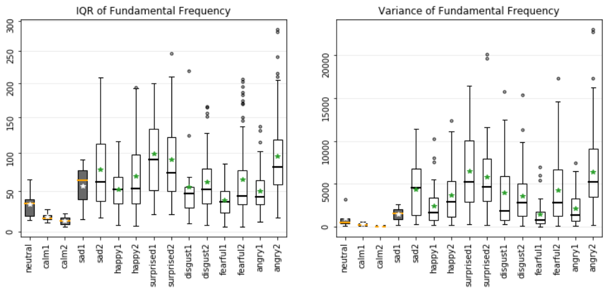
5 VOICE COACHING APPLICATION

Our voice coaching application is *Coach Ana*. See Figure 10. Coach Ana integrates all the previously mentioned voice analysis into a single mobile service which provides live voice coaching and reflective speech analysis. It seeks to help individuals match their speed, intensity, and intonation of their speech to their speaking goals.

Table 3: Pairwise Kruskal Analysis of Interquartile Range of Zero-Crossing Rate

Emotions have (1) normal and (2) strong versions except neutral

		Neutral	Calm		Happy		Sad		Angry		Fearful		Disgust		Surprise	
			1	2	1	2	1	2	1	2	1	2	1	2	1	2
Neutral		—	**3.528**	**1.671**	17.497	93.564	**0.006**	15.268	100.606	134.947	41.948	118.838	43.570	80.395	31.939	90.116
Calm	1	**3.528**	—	**0.154**	32.067	102.122	**2.483**	26.663	107.573	131.905	54.503	119.585	55.681	89.279	44.160	97.643
	2	**1.671**	**0.154**	—	24.883	94.967	**1.205**	21.116	100.720	128.446	48.350	115.562	49.312	82.494	37.906	90.354
Happy	1	17.497	32.067	24.883	—	59.318	14.667	**0.154**	70.677	128.908	15.127	103.416	16.603	55.057	7.103	57.808
	2	93.564	102.122	94.967	59.318	—	81.071	37.979	**1.752**	50.618	7.694	26.572	7.934	2.811	20.123	**1.288**
Sad	1	**0.006**	**2.483**	**1.205**	14.667	81.071	—	13.075	89.241	124.705	37.558	107.985	37.828	73.385	28.273	80.347
	2	15.268	26.663	21.116	**0.154**	37.979	13.075	—	48.550	104.497	9.075	79.309	9.334	40.914	**3.407**	42.304
Angry	1	100.606	107.573	100.720	70.677	**1.752**	89.241	48.550	—	33.432	14.405	14.490	13.759	**0.376**	29.411	**0.006**
	2	134.947	131.905	128.446	128.908	50.618	124.705	104.497	33.432	—	64.769	**4.195**	63.571	17.532	86.933	24.639
Fearful	1	41.948	54.503	48.350	15.127	7.694	37.558	9.075	14.405	64.769	—	42.487	**0.002**	14.081	**1.911**	12.066
	2	118.838	119.585	115.562	103.416	26.572	107.985	79.309	14.490	**4.195**	42.487	—	41.862	**6.238**	64.301	10.822
Disgust	1	43.570	55.681	49.312	16.603	7.934	37.828	9.334	13.759	63.571	**0.002**	41.862	—	13.090	**2.184**	11.952
	2	80.395	89.279	82.494	55.057	**2.811**	73.385	40.914	**0.376**	17.532	14.081	**6.238**	13.090	—	25.410	**0.138**
Surprised	1	31.939	44.160	37.906	7.103	20.123	28.273	**3.407**	29.411	86.933	**1.911**	64.301	**2.184**	25.410	—	23.660
	2	90.116	97.643	90.354	57.808	**1.288**	80.347	42.304	**0.006**	24.639	12.066	10.822	11.952	**0.138**	23.660	—



(a) IQR (b) Variance

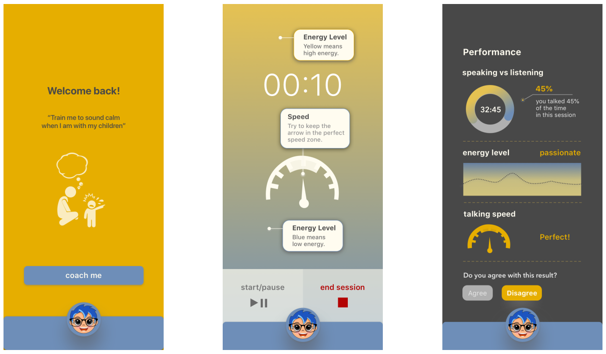
Figure 9: Variation Measures of Fundamental Frequency (F0 in Hz, shading of same groups from ZCR)

The application is designed for speech practice and for providing instantaneous feedback for real world speaking scenarios.

Different goals, audiences, and contexts for speaking will change what the appropriate speed intensity and intonation should be. Coach Ana particularizes the voice coaching for four different speech intentions, four different classes of audiences, and six speech environments, to give individuals personalized coaching for the specific kind of speaking they wish to improve. Including all combinations there are 96 speaking contexts. To find the target speaking affect, for example, "ideal speaking rate for teaching an elementary class", we collected findings from a diverse set of analyses and research relating to factors impacting perception [15, 18, 19, 21, 32, 33, 39],

Once a user has selected their speech context options, they can enter live coaching mode. In live coaching mode, the user is served a dashboard which displays their current speaking speed, intensity, and intonation; alongside goal values for each of these measures that they are to strive for during the speech. Speech metrics are updated every five seconds.

Aside from live coaching, Coach Ana also allows users to look back on their past sessions to get in depth analysis about a particular session or to visualize trends and track their long-term progress. See Figure 10(c).



(a) audience selection (b) live coaching (c) result presentation

Figure 10: Screenshots of the application.

Coach Ana has been on TestFlight since September 2019 for beta testing on iOS and watchOS devices, and is scheduled to release to the public in March 2020. The computation of acoustic measures is device independent and an Android version of the application is planned for after the iOS release.

6 CONCLUSIONS AND FUTURE WORK

This paper presents a framework for computing the acoustic measures of a sound or voice file, and for conducting experiments to find the relationships between the acoustic measures to auditory measures. An application is built for smartphones that provides real-time voice coaching to the speaker based on auditory measures inferred from acoustic measures. All computations are carried out on the users' mobile device without transmitting the contents of the speech utterance or sound signals off of the users' device.

We identified some ideas to experiment with for augmenting and improving the computations of the acoustic measures, including:

(1) improve voice activity detection with the state of the art LSTM-based VAD [37]
(2) compute loudness independently of speaker distance using speaker distance estimation [14]

(3) research and improve loudness measure for our application [23]
(4) discover more precise measure of intonation by exploring combinations of vocal features

Building on top of the acoustic measures we developed in this paper, more auditory measures can be further investigated and developed into the app to provide additional voice coaching functionality. Deeper experiments and analyses can be conducted on auditory properties such as:

- perceived speaker emotion
- elicited audience emotion
- intonation modulation

Another substantive improvement to our application Coach Ana is to conduct the acoustic analysis on continuous stream rather than once every 5 seconds. Implementing this involves decreasing the time between analysis while maintaining a time window sufficiently large for STM computation to be coherently analyzed. With some optimization, we conjecture that computing every second is possible using a rolling 10 second window of audio signal.

REFERENCES

[1] Vered Aharonson, Eran Aharonson, Katia Levi, Aviv Sotzianu, Ofer Amir, and Ovadia-Blechman Zehava. 2017. A Real-Time Phoneme Counting Algorithm and Application for Speech Rate Monitoring. *Journal of Fluency Disorders* 51 (01 2017). https://doi.org/10.1016/j.jfludis.2017.01.001

[2] Moataz M. H. El Ayadi, Mohamed S. Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition* 44 (2011), 572–587.

[3] Paul Boersma. 1993. ACCURATE SHORT-TERM ANALYSIS OF THE FUNDAMENTAL FREQUENCY AND THE HARMONICS-TO-NOISE RATIO OF A SAMPLED SOUND. In *Proceedings, Institute of Phonetic Science.* University of Amsterdam, 97 – 110.

[4] Paul Boersma and David Weenink. 2020. Praat: doing phonetics by computer [Computer program]. http://www.praat.org/

[5] Alain de Cheveigné and Hideki Kawahara. 2002. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America* 111 4 (2002), 1917–30.

[6] David Dean, Sridha Sridharan, Robert Vogt, and Michael Mason. 2010. The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms. In *Proceedings of 11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, K Hirose, S Nakamura, and T Kaboyashi (Eds.). 3110–3113.

[7] Laurence Devillers and Laurence Vidrascu. 2006. Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. In *INTERSPEECH 2006.*

[8] Laurence Devillers and Laurence Vidrascu. 2006. Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs.. In *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing.* 801 – 804.

[9] Dong Wang, Lie Lu, and Hong-Jiang Zhang. 2003. Speech segmentation without speech recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, Vol. 1. I–I. https://doi.org/10.1109/ICASSP.2003.1198819

[10] Sorin Dusan and Lawrence Rabiner. 2006. On the relation between maximum spectral transition positions and phone boundaries. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Vol. 2.

[11] Gustav Theodor Fechner, Edwin G Boring, and Davis H Howes. 1966. *Elements of Psychophysics: Transl. by Helmut E. Adler.* Holt, Rinehart and Winston.

[12] Sadaoki Furui. 1986. On the role of spectral transition for speech perception. *The Journal of the Acoustical Society of America* 80 (11 1986), 1016–25. https://doi.org/10.1121/1.393842

[13] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue. 1993. TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. https://catalog.ldc.upenn.edu/LDC93S1

[14] E. Georganti, T. May, S. van de Par, A. Harma, and J. Mourjopoulos. 2011. Speaker Distance Detection Using a Single Microphone. *IEEE Transactions on Audio, Speech, and Language Processing* 19, 7 (Sep. 2011), 1949–1961. https://doi.org/10.1109/TASL.2011.2104953

[15] Steven Greenberg, Hannah Carvey, Leah Hitchcock, and Shuangyu Chang. 2003. Temporal properties of spontaneous speech - a syllable-centric perspective. *J. Phonetics* 31 (2003), 465–485.

[16] Daniel Hirst and Albert di Cristo. 1999. A survey of intonation systems. In *Intonation Systems, A Survey of Twenty Languages*, Daniel Hirst and Albert di Cristo (Eds.). Cambridge University Press, Chapter 1, 1–44.

[17] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, and Raj Reddy. 2001. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development* (1st ed.). Prentice Hall PTR, USA.

[18] Judd Humpherys. 2012. Your Speech Patterns Affect Sales Performance. https://ezinearticles.com/?Your-Speech-Patterns-Affect-Sales-Performance&id=7306149

[19] Alexei Kapterev. 2011. *Presentation Secrets: Do What You Never Thought Possible with Your Presentations* (1st ed.). Wiley, USA.

[20] William H. Kruskal. 1952. A Nonparametric test for the Several Sample Problem. *Ann. Math. Statist.* 23, 4 (1952), 525–540.

[21] David Lewis and G. Riley Mills. 2012. *The Pin Drop Principle* (1st ed.). Jossey-Bass, USA.

[22] Steven R. Livingstone and Frank A. Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE* 13, 5 (05 2018), 1–35. https://doi.org/10.1371/journal.pone.0196391

[23] Lawrence Marks and Mary Florentine. 2010. *Measurement of Loudness, Part I: Methods, Problems, and Pitfalls.* 17–56. https://doi.org/10.1007/978-1-4419-6712-1_2

[24] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference.* 18 – 25.

[25] Vikramjit Mitra and Elizabeth Shriberg. 2015. Effects of feature type, learning algorithm and speaking style for depression detection from speech. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015.* 4774–4778. https://doi.org/10.1109/ICASSP.2015.7178877

[26] Nelson Morgan, Eric Fosler-Lussier, and Nikki Mirghafori. 1997. Speech Recognition Using On-Line Estimation Of Speaking Rate. In *Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997*, Vol. 4.

[27] Gregor Pirker, Michael Wohlmayr, Stefan Petrik, and Franz Pernkopf. 2011. A Pitch Tracking Corpus with Evaluation on Multipitch Tracking Scenario. In *INTERSPEECH 2011.* 1509–1512.

[28] Tim Polzehl, S. Möller, and Florian Metze. 2010. Automatically Assessing Personality from Speech. In *IEEE Fourth International Conference on Semantic Computing (ICSC).* 134 – 140. https://doi.org/10.1109/ICSC.2010.41

[29] The WebRTC project authors. 2011 (accessed Novemver, 2019). The WebRTC project. https://webrtc.org/

[30] Thomas F. Quatieri. 2001. *Discrete-Time Speech Signal Processing: Principles and Practice* (1st ed.). Prentice Hall, USA.

[31] Mirco Ravanelli, Titouan Parcollet, and Yoshua Bengio. 2018. The Pytorch-kaldi Speech Recognition Toolkit. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), 6465–6469.

[32] Emma Rodero. 2012. A comparative analysis of speech rate and perception in radio bulletins. *Text and Talk* 32 (05 2012), 391–411. https://doi.org/10.1515/text-2012-0019

[33] Bruce P. Ryan. 2000. Speaking rate, conversational speech acts, interruption, and linguistic complexity of 20 pre-school stuttering and non-stuttering children and their mothers. *Clinical Linguistics & Phonetics* 14, 1 (2000), 25–51. https://doi.org/10.1080/026992000298931 arXiv:https://doi.org/10.1080/026992000298931 PMID: 22091696.

[34] B. Schuller. 2011. Voice and speech analysis in search of states and traits. In *Computer Analysis of Human Behavior.* Springer, 227 – 253.

[35] Björn W. Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian A. Müller, and Shrikanth S. Narayanan. 2010. The INTERSPEECH 2010 paralinguistic challenge. In *INTERSPEECH 2010.*

[36] Piotr Sorokowski, David Puts, Janie Johnson, Olga Żółkiewicz, Agnieszka Sorokowska, Marta Kowal, Basia Borkowska, and Katarzyna Pisanski. 2019. Voice of Authority: Professionals Lower Their Vocal Frequencies When Giving Expert Advice. *Journal of Nonverbal Behavior* (05 2019). https://doi.org/10.1007/s10919-019-00307-0

[37] S. Tong, H. Gu, and K. Yu. 2016. A comparative study of robustness of deep learning approaches for VAD. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 5695–5699. https://doi.org/10.1109/ICASSP.2016.7472768

[38] Benjamin Weiss and Felix Burkhardt. 2012. Is 'not bad' good enough? Aspects of unknown voices' likability. In *INTERSPEECH 2012.*

[39] Jiahong Yuan, Mark Liberman, and Christopher Cieri. 2006. Towards an Integrated Understanding of Speaking Rate in Conversation. In *Proceedings of INTERSPEECH 2006.*